分布式数据库技术的研究与实现

杨东[1]，谢菲[2]，杨晓刚[2]，何遵文[1]
（1.北京理工大学&引跑科技联合大数据实验室，北京，100081；
2.新华社713实验室，北京，100803）

摘　要：分布式数据库技术是数据库技术与计算机网络技术[1]相结合的产物,主要技术涉及并行计算、分布策略、数据分片、查询优化以及分布式数据库系统的并发控制[2]、事务处理与恢复技术等。本文旨在通过对分布式数据库关键技术的研究，找到一种新型的、适应多种复杂应用场景的通用型分布式数据库模型，以应对日益凸显的大数据变革带来的挑战。

关键词：海量数据、分布式数据库、并行计算、数据分片
中图分类号： TP331　　**文献标识码：** A　　**文章编号：** 2095-8595 (2015) 01-087-08
电子科学技术 URL: http//www.china-est.com.cn　DOI: 10.16453/j.issn.2095-8595.2015.01.017

Research and Implementation of Distributed Database Technology

Fei Xie, Xiaogang Yang, Zunwen He, Sudong Yang, Dong Yang
(*1.Beijing Institute of Technology &Intple Technology Big Data Joint Laboratory, Beijing,* 100081, *China;*
2. Xinhua News Agency 713 Laboratory, Beijing, 100803, *China*)

Abstract: Distributed Database technology is progeny combining database technology with computer network technology. It involves parallel computing, distribution strategy ,data slice ,query optimization and distributed database concurrency control, transaction processing and recovery, etc. This paper aims to study the key technology of the distributed database, to find a new,general-purpose distributed database model to adapt to a variety of complex application scenarios ,with large data change in response to increasingly challenge.

Key words: Massive Data; Distributed Database; Parallel Computing; Data Slice

1 传统数据库面临的问题与挑战

近年来，有关大数据分析的讨论正在愈演愈烈。甚至出现了爆炸性增长的趋势，为什么大数据管理和分析突然之间受到了如此关注？一方面是由于移动互联网和移动智能终端的普及发展，数据信息正以每年40%的速度增长，造成数据量庞大；同时，数据种类呈多样性，文本、图片、视频等结构化和非结构化数据共存；另一方面也要求实时交互性强；最重要的是大数据蕴含了巨大的商业价值。

随之而来，对于海量异构数据[11]的灵活管理及并行处理提出了更高的要求，基于传统架构的数据库技术[4]在高扩展性、高性能等方面都已难以应对，需要基于新型分布式数据库处理技术[3],[7]解决日益凸显的大数据管理及计算难题。

2 分布式数据库应具备的特性

- MPP无共享架构

MPP无共享架构是当前分布式数据库最优化的I/O

处理架构，所有的节点同时进行并行处理，节点之间完全无共享，无I/O冲突；MPP无共享架构增加节点实现线性扩展，增加节点可线性增加存储、查询和加载性能。

• 高可用性

分布式数据库需要通过多份数据冗余机制避免单点故障，同时支持在线故障节点数据重建，支持在线不同粒度数据迁移、备份与恢复来体现高可用性，保证面对各种异常时可以提供正常服务的能力。

• 高性能

分布式数据库的高可用性表现在的吞吐能力以及系统的响应时间，其查询、写入性能应随分布式集群规模的扩展准线性增长。

• 自动数据分片

分片是指将数据拆分，分散到不同的数据库实例上进行“负载分流”的做法。 分布式数据库系统通过片键（shard key）定义进行数据分片，支持递增片键（连续、不均匀、写入集中），随机片键（不连续、均匀、写入分散、分流较好）。

• 智能水平扩充

通过自动数据分片、复制[5],[13]以及容错等机制，能够将分布式数据库系统部署到多台服务器，使得分布式数据库集群[14]可以根据业务的特点进行弹性扩容。

3 分布式数据库关键技术的研究与实现

3.1 技术架构设计

为了满足分布式数据库在高可用、高性能、高扩展方面的需求，本文提出了一种新型分布式数据库处理架构设计理念（见图1所示），其架构包括分布式数据库引擎和分布式数据存储节点两个部分。分布式数据库引擎是系统核心，其负责SQL解析、优化、路由[10]、分发、合并等操作，同时将底层的众多存储节点管理起来；分布式存储节点使用关系型数据库，主要负责数据存储、处理及同步[6],[8]。在实际应用中可灵活构建不同规模的数据库集群，通过将业务数据分片到不同的数据库存储节点中，极大地降低了普通数据库面对海量数据时的压力；通过将用户的SQL请求分发到各节点上执行，充分利用各节点的计算资源，从而使PC服务器集群达到小型机、中大型机的性能。

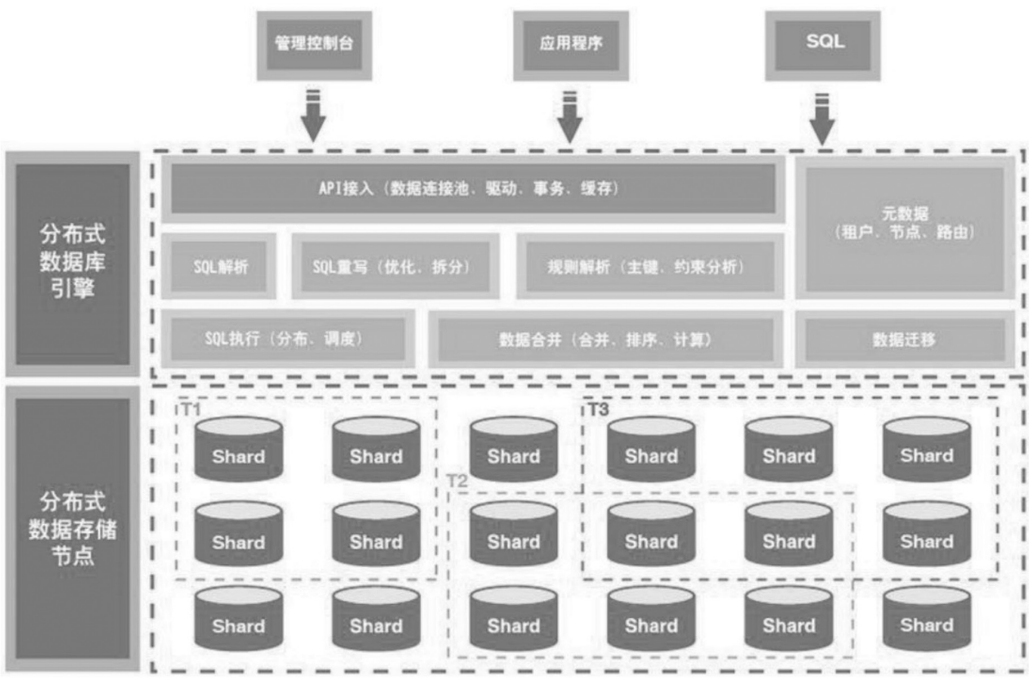

图1 分布式数据库处理架构模型

3.2 并行计算服务

分布式数据库依赖于分布式并行计算服务[9]实现海量内存支持、计算任务多核多机支持、高度定制化的实时并行计算集群系统。并行计算服务包括节点发现、节点通信和节点路由协同等技术，充分发挥计

算潜力，实现弹性计算，全面支持大数据场景下复杂计算需求。分布式计算架构及集群对等架构分别如图2(a)、(b)所示。

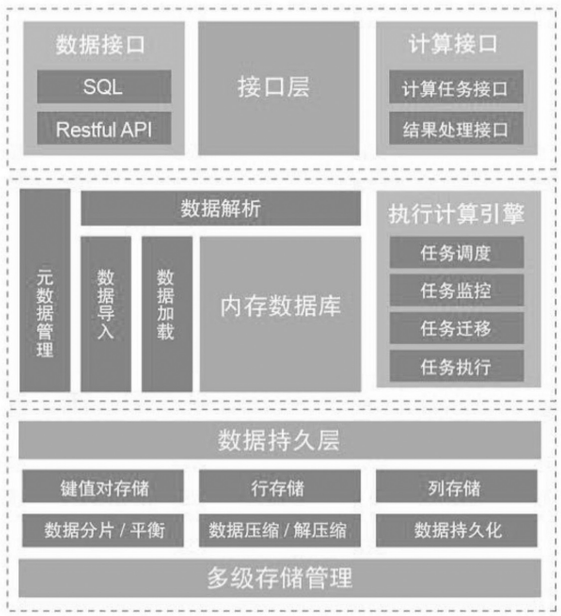

图2（a） 并行计算架构模型

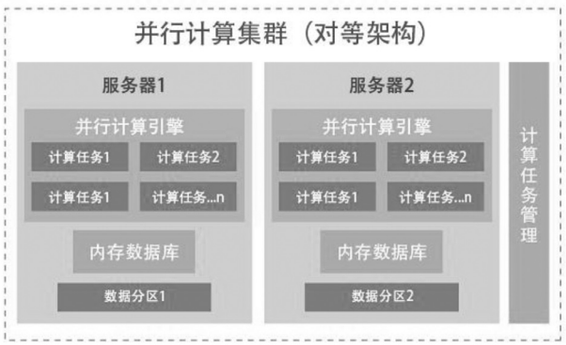

图2（b）并行计算集群对等架构

分布式数据库的并行处理主要体现在外部数据并行装载、并行备份恢复与并行查询处理三个方面。并行处理的主要应用场景是大数据分析领域中的数据装载和查询，数据的并行装载主要是在采用外部表或者Web方式，通过数据库引擎层提供的Dataloader来实现。在ETL带宽足够的情况下，可以运行多个Data Loader程序并行装载各种文件，或者根据数据库引擎CPU核心数量运行Data Loader程序装载。

3.3 基于哈希优化算法的数据分片技术

哈希分布是目前最常用的数据分片[12]技术，即在数据库建表阶段，根据表的主键、外键约束、唯一键约束、表字段类型等因素选择、判断，将哈希值相同的数据记录在同一个数据节点。这里我们通过优化哈希算法使得数据的分布更为均匀、合理。

CREATE TABLE … DISTRIBUTED BY (column [,…]），哈希值相同的记录在同一个数据节点，如图3所示:

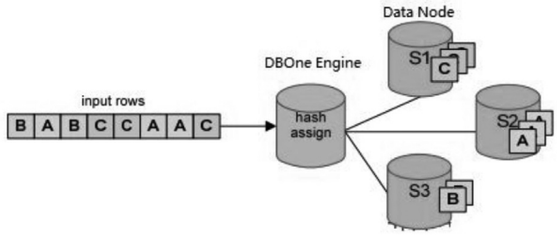

图3 数据分布式方法-哈希分布

数据的一致性保证通过以下操作完成:

首先，求出每个服务节点的hash，并将其配置到一个0 到2^{32}的圆环（continuum）区间上。

其次，使用同样的方法求出你所需要存储的key的hash值，也将其配置到这个圆环（continuum）上。

最后，从数据映射到的位置开始顺时针查找，将数据保存到找到的第一个服务节点上。如果超过2^{32}仍然找不到服务节点，就会保存到第一个memcached服务节点上。如图4所示:

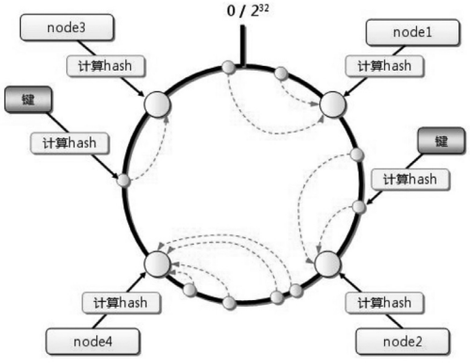

图4 数据的一致性保证

当增加服务节点时,数据图例如下:

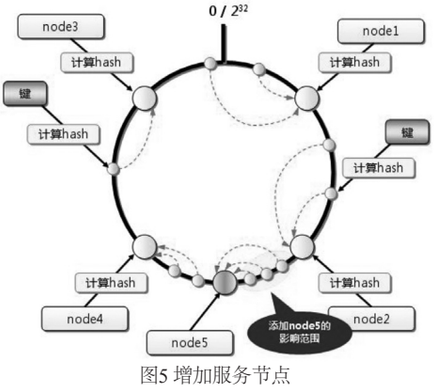

图5 增加服务节点

其他：只有在圆环上增加服务节点的位置为逆时针方向的第一个服务节点上的键会受到影响。

数据分布实验：选取9个不同数据量的样本表，选择10到64个分片，通过哈希优化算法进行数据分片，从实验结果如表1可以看出，数据分布误差率基本控制在5%以内，达到了预期。

表1 数据分布实现数据

表样本	总数据量（亿）	路由字段	分片算法	总分片数	每分片数据条数（万）	数据分布偏差率
T_3A_CDR_LIST	1.2	s_mdn	hash	40	287-312	4.33%
huifu	10	id	hash	40	2423-2582	3.28%
t_item	4	id	hash	32	1217-1298	3.84%
CK10_CFMX	20	id	hash	64	3053-3185	2.30%
ENTRY_LIST	10	entry_id	hash	36	2745-2860	2.14%
ENTRY_WORKFLOW	50	entry_id	hash	36	13736-14131	1.09%
dwb_rtl_sls_retrn_line_item_d	6	, trx_line_item_	hash	36	1612-1717	3.24%
adputlog	2	adid	hash	20	961-1038	3.90%
wf_task_log	0.8	bind_id	hash	10	769-833	4.13%

3.4 分布式数据库的设计与实现

在关键技术研究的基础上，设计并实现分布式数据库系统原型，系统原型功能包括：数据模型模块、数据库管理模块、状态监控模块、备份与恢复模块及日志与审计模块。其中的数据模型模块（见图6）包括应用管理、表管理（见图7）、视图管理、序列管理、存储视图管理及自定义类型管理等功能；数据库管理模块（见图8）包括节点管理、实例管理、数据库管理及分发管理等功能；状态监控模块（见图9）包括监控总览、服务器状态监控、数据库状态监控、连接池状态监控及引擎状态监控等功能；备份与恢复模块（见图10）包括系统备份及备份策略等功能；日志与审计模块（见图11）包括系统日志、审计日志及SQL执行计划等功能。管理界面和数据分片界面分别如图所示。

图6 模型管理原型界面

图7 表管理数据分片原型界面

图8 数据库管理原型界面

图9 状态监控原型界面

图10 备份与恢复原型界面

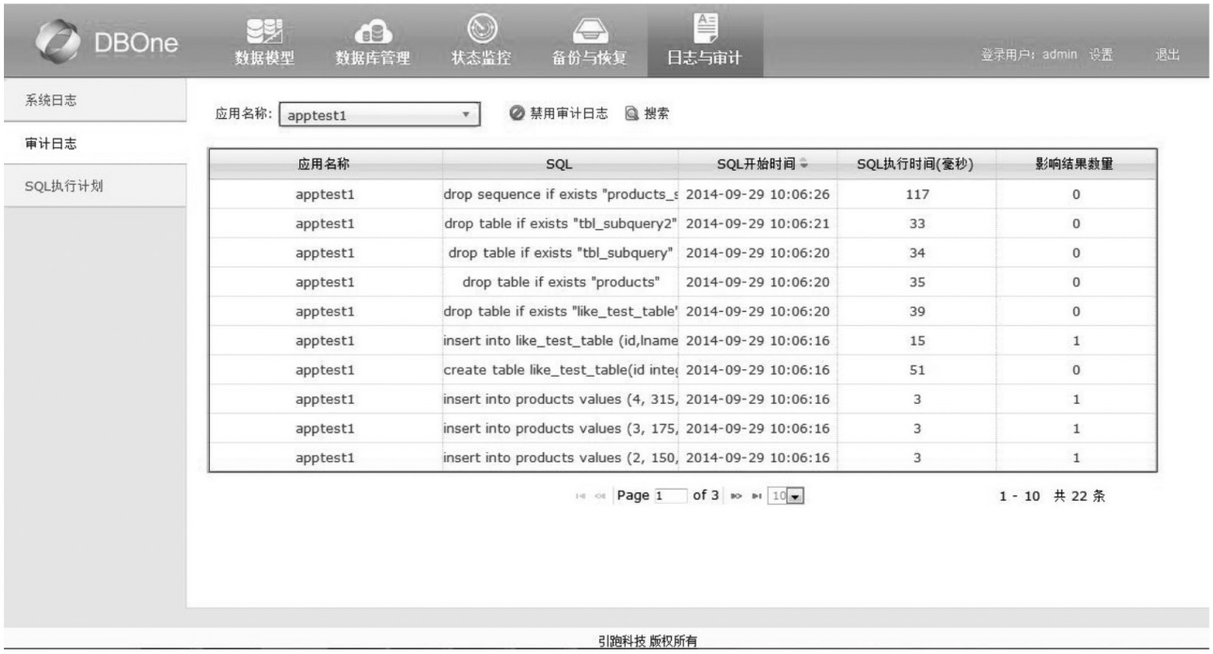

图11 日志与审计原型界面

4 验证及结论

为了验证本文提出的分布式数据库技术的可行性和实际效果，选取新媒体行业作为应用实践场景进行验证。新媒体应用架构如图12所示。新媒体行业自身特点决定了其业务数据的多样性且体量巨大，如何为日益增加的海量异构数据和多元化的信息载体提供强有力的数据管理平台已成为新媒体技术的发展面临的首要问题。

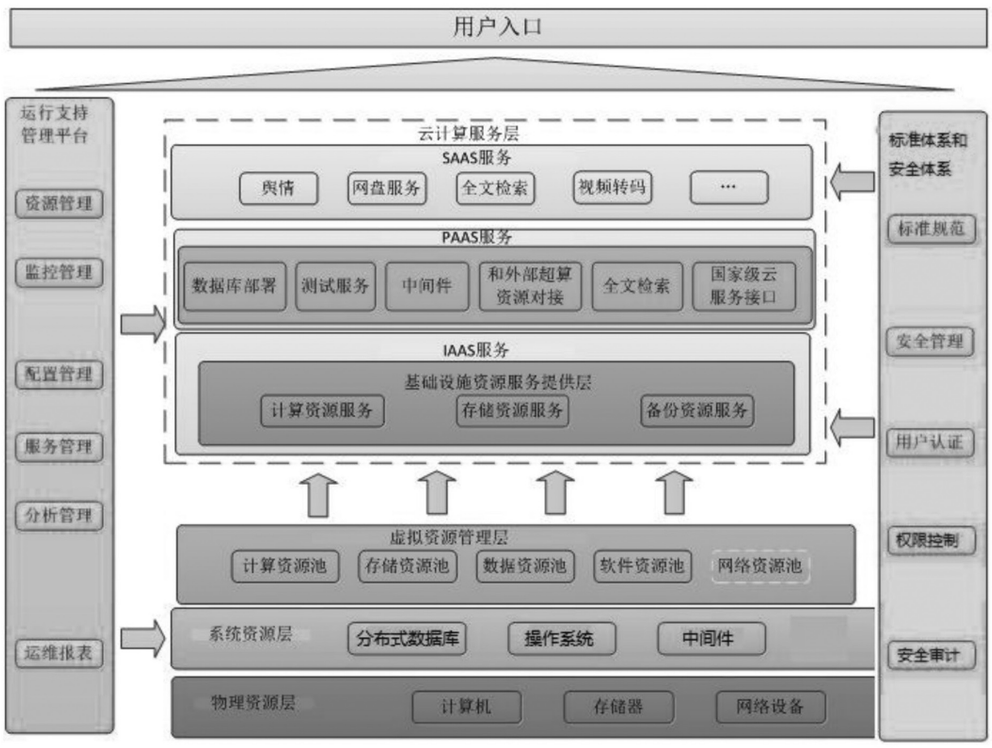

图12 新媒体应用架构

通过验证，分布式数据库处理架构实现了对这些业务所产生的海量结构化数据和半结构化数据等异构数据类型的统一、高效管理；分布式数据库引擎数据中的数据装载程序（Dataloader）能够以420MB/s装载text格式的文件和300MB/s装载CSV格式文件；通过分布式的数据存储节点实现业务数据的规模存储和按需节点增加，保证了在不中断现有业务应用的情况下，实现海量数据存储节点的弹性扩展，同时平台的性能也随节点增加准线性提升，从而有效应对新媒体海量数据处理的压力。

参考文献

[1] Quantum Information Scienceand TechnologyQuIST program ver.2.0[J].Defense Advanced Research ProjectsAgency DARPA,2004,4.

[2] Karl Johan.Wittenmark.Computer-Controlled Systems (3rd ed.).Prentice Hall.1997.

[3] 李霖,周兴铭.分布式数据库研究的新方向[J].计算机应用与软件. 2000(06).

[4] 姚文琳,王存刚,刘世栋,仇利克. 基于Oracle的分布式数据库设计与技术[J]. 计算机工程. 2006(20).

[5] 盖九宇,张忠能,肖鹤.分布式数据库数据复制技术的分析与应用[J].计算机应用与软件. 2005(07).

[6] 赵应钢.异构分布式数据库数据同步系统设计与实现[D].华中科技大学 2007.

[7] 魏少华.基于WEB的分布式数据库系统的研究与设计[D].西北工业大学 2007.

[8] 张雄. 分布式数据库数据同步的研究与应用[D].华中科技大学 2006.

[9] 戴炳荣,宋俊典,钱俊玲. 云计算环境下海量分布式数据处理协同机制的研究[J]. 计算机应用与软件. 2013(01).

[10] 刘威.分布式数据库及其技术[J].长春大学学报. 2000(01) .

[11] 胡彬华,李晓,梁剑.异构分布式数据库系统集成的研究与实现[J].计算机应用研究.2002(10).

[12] 杨艺.分布式数据库中数据分配方法的研究[D]. 重庆大学 2004.

[13] 王婉菲,张志浩.分布式数据库系统的复制机制及应用[J].计算机工程与科学. 2003(01).

[14] 王婉菲,王欣,张志浩.数据库集群系统的研究与实施[J].微型电脑应用.2003(10).

作者简介：

杨东(1981-)，研究员（方案总监），硕士，北京理工大学&引跑科技联合大数据实验室（上海引跑信息科技有限公司），主要研究方向是分布式数据库数据分布及并行处理技术及应用实现。

谢菲(1981-)，高级工程师，硕士，新华社713实验室，主要研究大数据与智能信息处理技术在新媒体行业的应用与实践。

杨晓刚(1983-)，工程师，博士，新华社713实验室，主要研究大数据与智能信息处理技术在新媒体行业的应用与实践。

何遵文(1964-)，副教授，博士，北京理工大学信息与电子学院，主要研究方向无线传感网。